Penalized regression combining the L1 norm and a correlation based penalty.

Mohammed El Anbari, Abdallah Mkhadri

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Penalized regression combining the L_1 norm and a correlation based penalty

Mohammed El Anbari — Abdallah Mkhadri

N° 6746

December 2008

Thème COG

apport de recherche

ISSN 0249-6399 ISRN INRIA/RR--6746--FR+ENG

Penalized regression combining the L_1 norm and a correlation based penalty

Mohammed El Anbari*†, Abdallah Mkhadri† †

Thème COG — Systèmes cognitifs
Équipe-Projet Select et Cadi Ayyad University

Rapport de recherche n° 6746 — December 2008 — 32 pages

Abstract: Variable selection in linear regression can be challenging, particularly in situations where a large number of predictors is available with possibly high correlations, such as gene expression data. In this paper we propose a new method called the elastic corr-net to simultaneously select variables and encourage a grouping effect where strongly correlated predictors tend to be in or out of the model together. The method is based on penalized least squares with a penalty function that, like the Lasso penalty, shrinks some coefficients to exactly zero. Additionally, this penalty contains a term which explicitly links strength of penalization to the correlation between predictors. A detailed simulation study in small and high dimensional settings is performed, which illustrates the advantages of our approach in relation to several other possible methods. Finally, we apply the methodology to three real data sets. The key contribution of the elastic corr-net is the identification of setting where the elastic net fails to product good results: in terms of prediction accuracy and estimation error, our empirical study suggests that the elastic corr-net is more adapted than the elastic-net to situations where $p \leq n$ (the number of variables is less or equal to the sample size). if $p \gg n$, our method remains competitive and also allows the selection of more than n variables in a new way.

Key-words: Variable selection; high dimensional setting; elastic net; grouping effect; correlation based penalty.

The first author is partly supported by the project Maroc-STIC.

* INRIA Futurs, Projet select, Université Paris-Sud 11
† Libma laboratory, cadi ayyad university, Morroco

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

La régression pénalisée combinant la norme L_1 et une pénalité tenant compte des corrélations entre les variables

Résumé : La sélection de variables peut être difficile, en particulier dans les situations où un grand nombre de variables explicatives est disponible, avec la présence possible de corrélations élevées comme dans le cas des données d'expression génétique. Dans cet article, nous proposons une nouvelle méthode de régression linéaire pénalisée, appelée l'*elastic corr-net*, pour simultanément estimer les paramètres inconnus et sélectionner les variables importantes. De plus, elle encourage un effet de groupe: les variables fortement corrélées ont tendance à être toutes incluses ou toutes exclues du modèle. La méthode est fondée sur les moindres carrés pénalisés avec une pénalité qui, comme la pénalité L_1, rétrécit certains coefficients exactement vers zéro. En outre, cette pénalité contient un terme qui lie explicitement la force de pénalisation à la corrélation entre les variables explicatives. Pour montrer les avantages de notre approche par rapport aux méthodes les plus concurrentes, une étude détaillée de simulation est réalisée en moyenne et grande dimension. Enfin, nous appliquons la méthodologie à trois ensembles de données réelles. Le résultat principal de notre méthode est l'identification du cadre où l'elastic-net est moins performante : en effet, en termes des erreurs de prédiction et d'estimation, notre méthode paraît plus adaptée aux situations du type $p \leq n$ (le nombre de variables est inférieure à la taille de l'échantillon). Si $p \gg n$, notre méthode reste compétive et elle permet aussi de sélectionner plus que n variables.

Mots-clés : Sélection de variables; grandes dimensions; elastic-net; effet groupement; pénalité de corrélation.

1 Introduction

We consider the standard linear regression model

$$y = \beta_0 + \mathbf{x}^T\beta + \varepsilon, \tag{1.1}$$

where $\mathbf{x} = (x_1, ..., x_p)^T$ is the vector of covariates and ε is a noise variable with $\mathbb{E}(\varepsilon) = 0$. The vector of predicted responses is given by

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1\mathbf{x}_1 + ... + \hat{\beta}_p\mathbf{x}_p. \tag{1.2}$$

Interest focuses on finding the vector $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ that is best under a given criterion, such as prediction accuracy.

Ordinary least squares (OLS) computes $\hat{\beta}$ by minimizing the residual sum of squares. Despite its simplicity and unbiasedness, the OLS estimator is, however, not always satisfactory in both prediction and interpretation.

Most of the alternatives can be categorized into one of two groups. The first set of approaches uses some form of regularization on the regression coefficients to accept a small increase of the bias for a possibly decrease in of the variance. Among these there is the ridge regression (RR) (Hoerel and Kennard, 1970), which minimizes the sum of residuals squares subject to a bound on the L_2 norm of the coefficients, or more recently the Correlation Penalty (CP) (Tutz and Ulbricht (2006)), which explicitly use the correlation between predictors in the L_2 norm penalty term. While these approaches often produce improvements in prediction accuracy, the final fit may be difficult to interpret because all p variables remain in the model. The second set of approaches begins by performing variable selection, i.e. determining which $\hat{\beta}_j = 0$ for some j. By implementing OLS on the reduced number of variables one can often gain increased prediction accuracy as well as a more easily interpretable model.

In the last decade interest has focused on an alternative class of methods which implement both the variable selection and the coefficient shrinkage in a single procedure. The most well known of these procedures is the Lasso (Tibshirani, 1996; Chen et al., 1998). The Lasso uses an L_1 penalty on the coefficients, which has the effect of automatically performing variable selection by setting certain coefficients to zero and shrinking the remainder. This method was made particularly appealing by the advent of the LARS algorithm (Efron et al., 2004) which provided a highly efficient mean to simultaneously produce the set of Lasso fits for all values of the tuning parameter.

Although it is a highly successful technique, it has two drawbacks:

i) In $p > n$ case, the Lasso can select at most n variables, this can be a limiting feature for a variable selection method.

ii) When there are several highly correlated input variables in the data set, all relevant to the output variable, the L_1-norm penalty tends to pick only one or few of them

and shrinks the rest to 0. For example, in microarray analysis, expression levels for genes that share a biological pathway are usually highly correlated, and these genes all contribute to the biological process, but the L_1-norm penalty usually selects only one gene from the group and does not care which one is selected. An ideal method should be able to eliminate trivial genes, and automatically include the whole group of relevant genes.

Recently, Zou and Hastie (2005) proposed the elastic-net as an alternative procedure which handles the deficiencies of Lasso and ridge regression by combining L_1 and L_2 penalties. The elastic-net has the ability to include group of variables which are highly correlated. In the same spirit, Bondell and Reich (2007) proposed a new method called Oscar to simultaneously select variables and perform supervised clustering in the context of linear regression. The technique is based on the penalized least squares with a penalty function combining the L_1 and the pairwise L_∞ norms. Moreover, the computation of the Oscar estimates are based on a sequential quadratic programming algorithm which can be slow for large p. While the elastic-net seems to be slightly less reliable in the presence of positively and negatively correlated variables. However, it is remarked that the elastic-net is particularly adapted to high dimensional setting ($n << p$)

On the other hand, the correlation based estimator CP does shrinkage but not variable selection. In order to obtain the grouping effect of CP in combination with variable selection, Tutz and Ulbricht (2006) proposed boosting procedure (called blockwise boosting) which updates at each step the coefficient of more than one variable. But, in practical implementation, the step length factor and the stopping number of iterations have to be determined. This sometimes may be difficult, and can affect the sparsity of the solution as well as the speed of convergence of the algorithm.

In this paper, we propose an alternative regularization procedure based on the penalized least squares for variable selection in linear regression problem, which combines the L_1 norm and CP penalties. We call it the *elastic corr-net*. Similar to the elastic-net method, the elastic corr-net performs automatic variable selection and parameter estimation where highly correlated variables are able to be selected (or removed) together. Additionally, the CP penalty contains a term which explicitly links strength of penalization to the correlation between variables. In contrast to the elastic-net, our approach seems to be adapted to both small and high dimensional settings : $p \leq n$ and $n << p$.

The remainder of this paper is organized as follow. In Section 2, we formulate the elastic corr-net as a constrained least squares problem using a novel *elastic corr-net penalty*. We discuss the grouping effect that is caused by the elastic corr-net penalty. Computational issues, including choosing the tuning parameters, are discussed in Section 3. Section 4 is devoted to numerical experimentations on simulated data which show, that the elastic corr-net compares favorably to the existing shrinkage and variable selection techniques in terms of both prediction error and identification of relevant variables. Finally, the elastic corr-net is applied to the body fat, the NewsUS data and the

Ionosphere data sets in Section 5. We end the paper with a brief discussion in section 6.

2 The elastic corr-net

In this section, we present the elements of the elastic corr-net algorithm.

2.1 The criterion

Suppose that the data set has n observations with p predictors. Let $\mathbf{y} = (y_1, ..., y_n)^T$ be the response and $\mathbf{X} = (\mathbf{x}_1|...|\mathbf{x}_p)$ be the model matrix, where $\mathbf{x}_j = (x_{1j}, ..., x_{nj})^T, j = 1, ..., p$, are the predictors. It is assumed that the response is centered and the predictors are standardized,

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad \text{for } j = 1, ..., p. \tag{2.3}$$

The elastic corr-net criterion solves

$$\min_\beta \textstyle\sum_{k=1}^{n} (y_k - \mathbf{x}_k^T \beta)^2$$

$$\text{subject to} \quad \begin{cases} \|\beta\|_1 \leq s_1 \\ P_c(\beta) \leq s_2 \end{cases} \tag{2.4}$$

where

$$P_c(\beta) = \sum_{j=1}^{p-1} \sum_{j>i} \left\{ \frac{(\beta_i - \beta_j)^2}{1 - \rho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1 + \rho_{ij}} \right\}, \tag{2.5}$$

$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$, s_1 and s_2 are nonnegative values.

The first penalty encourages sparsity in the coefficients. The term ρ_{ij} denotes the (empirical) correlation between the ith and the jth predictor. It is designed in a way so that for strong positive correlation ($\rho_{ij} \approx 1$), the first term becomes dominant having the effect that estimates for β_i and β_j are similar ($\hat{\beta}_i \approx \hat{\beta}_j$). For strong negative correlation ($\rho_{ij} \approx -1$), the second term becomes dominant and $\hat{\beta}_i$ will be close $-\hat{\beta}_j$. The effect is grouping; highly correlated variables lead to comparable values of estimates ($|\hat{\beta}_i| \approx |\hat{\beta}_j|$).

The penalty $P_c(\beta)$ was introduced by Tutz and Ulbricht (2006) as an alternative to the L_2 norm in ridge regression method, and it results on a correlation based penalty method (CP hereafter).

A nice feature of the penalty (2.5) is that it may be written as a simple quadratic form:

$$P_c(\beta) = \beta^T \mathbf{W} \beta,$$

where $\mathbf{W} = (w_{ij})_{1\leq i,j\leq p}$ is a matrix with general term, assuming that $\rho_{ij}^2 \neq 1$ for $i \neq j$,

$$w_{ij} = \begin{cases} 2\sum_{s\neq i} \frac{1}{1-\rho_{is}^2}, & i = j \\ -2\frac{\rho_{ij}}{1-\rho_{ij}^2}, & i \neq j \end{cases}$$

(for proof see Tutz and Ulbricht 2006).

The Lagrangian for the elastic corr-net is

$$L(\lambda_1, \lambda_2, \beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2 P_c(\beta), \tag{2.6}$$

for any fixed non negative λ_1 and λ_2.

The criterion (2.6) can be viewed as a penalized least squares criterion. Let $\alpha = \lambda_1/(\lambda_1 + \lambda_2)$; then estimating $\hat{\beta}$ is equivalent to the optimization problem

$$\hat{\beta} = \arg\min \|\mathbf{y} - \mathbf{X}\beta\|^2, \text{s.t } (1-\alpha)\|\beta\|_1 + \alpha P_c(\beta) \leq t \text{ for } t \geq 0. \tag{2.7}$$

We call the function $(1-\alpha)\|\beta\|_1 + \alpha P_c(\beta)$ the *elastic corr-net* penalty, it is a convex combination of the Lasso and correlation based penalty. The role of the L_1-norm penalty is to allow for variable selection, and the role of the P_c penalty is to get groups of correlated variables selected together. The elastic-corr net penalty is singular at 0 and it is strictly convex for all $\alpha \in [0, 1)$, thus having the characteristics of both the Lasso and CP regression methods. Figure 1 shows the contour plots of the penalty for three amounts of positive and negative correlations.

2.2 Estimation

We now develop a method to estimate the elastic corr-net model efficiently. It turns out that minimizing criterion (2.6) is equivalent to a Lasso-type optimization problem. This fact implies that the new method can enjoy the computational advantage of the Lasso.

Because $\mathbf{W}$ is a real symmetric positive-definite square matrix, it admits a Choleski decomposition: it exists an upper triangular matrix $\mathbf{L}$ so that

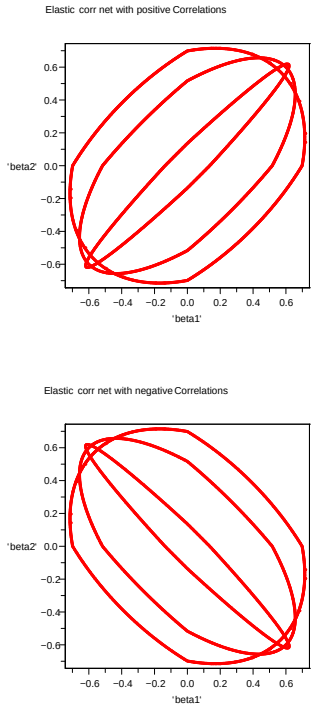
$$\mathbf{W} = \mathbf{L}\mathbf{L}^T. \tag{2.8}$$

The following Lemma is similar to the Lemma 1 in Zou and Hastie (2005), except that, we use the Choleski decomposition $\mathbf{L}$ instead of the identity matrix in the augmented covariates matrix $\mathbf{X}^*$.

Lemma 1 *Given* $(\mathbf{y}, \mathbf{X})$, (λ_1, λ_2), *and the Choleski factorization of* $\mathbf{W}$ *(2.8), define an augmented data set* $(\mathbf{y}^*, \mathbf{X}^*)$ *by*

$$\mathbf{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{L}^T \end{pmatrix}, \quad \mathbf{y}^*_{(n+p)} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

Figure 1: Top panel: Two-dimensional contour plots of $0.5\|\beta\|_1 + 0.5P_c(\beta) = 1$ for three amounts of positive correlation: $\rho = 0.5$, $\rho = 0.8$, and $\rho = 0.99$. Bottom panel: Two-dimensional contour plots of $0.5\|\beta\|_1 + 0.5P_c(\beta) = 1$ for three amounts of negative correlation: $\rho = -0.5$, $\rho = -0.8$, and $\rho = -0.99$.



Let $\gamma = \lambda_1/\sqrt{1+\lambda_2}$ *and* $\beta^* = \sqrt{1+\lambda_2}\beta$. *Then criterion (2.6) can be written as*

$$L(\gamma, \beta) = L(\gamma, \beta^*) = \|\mathbf{y}^* - \mathbf{X}^*\beta^*\|_2^2 + \gamma\|\beta^*\|_1.$$

Let

$$\hat{\beta}^* = \arg\min_{\beta^*} L\{(\gamma, \beta^*)\},$$

then the elastic corr-net estimates defined in (2.6) verifies

$$\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*.$$

Proof. We have

$$\|\mathbf{y}^* - \mathbf{X}^*\beta^*\|_2^2 = \mathbf{y}^{*T}\mathbf{y}^* - 2\mathbf{y}^{*T}\mathbf{X}^*\beta^* + \beta^{*T}\mathbf{X}^{*T}\mathbf{X}^*\beta^*.$$

From the identities

$$\begin{aligned}
\mathbf{X}^{*T}\mathbf{X}^* &= (1+\lambda_2)^{-1}\left[\mathbf{X}^T\sqrt{\lambda_2}\mathbf{L}\right]\begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{L}^T\end{bmatrix} \\
&= (1+\lambda_2)^{-1}\left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{L}\mathbf{L}^T\right) \\
&= \left(\frac{\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{W}}{1+\lambda_2}\right)
\end{aligned}$$

$$\mathbf{y}^{*T}\mathbf{X}^* = \frac{\mathbf{y}^T\mathbf{X}}{\sqrt{1+\lambda_2}},$$

$$\mathbf{y}^{*T}\mathbf{y}^* = \mathbf{y}^T\mathbf{y},$$

$$\gamma\|\beta^*\|_1 = \lambda_1\|\beta\|_1,$$

we get

$$\beta^{*T}\mathbf{X}^{*T}\mathbf{X}^*\beta^* = \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda_2\beta^T\mathbf{W}\beta.$$

Finally we have:

$$\begin{aligned}
\|\mathbf{y}^* - \mathbf{X}^*\beta^*\|_2^2 + \gamma\|\beta^*\|_1 &= \mathbf{y}^{*T}\mathbf{y}^* - 2\mathbf{y}^{*T}\mathbf{X}^*\beta^* + \beta^{*T}\mathbf{X}^{*T}\mathbf{X}^*\beta^* + \gamma\|\beta^*\|_1 \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda_2\beta^T\mathbf{W}\beta + \lambda_1\|\beta\|_1 \\
&= \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\beta^T\mathbf{W}\beta \\
&= L(\lambda_1, \lambda_2, \beta).
\end{aligned}$$

This completes the proof ■

Lemma 1 says that we can transform the elastic corr-net into an equivalent Lasso problem on augmented data. Note that the sample size in the augmented data is $n+p$ and $\mathbf{X}^*$ has rank p, which means that the new method can potentially select all p predictors in all situations. Lemma 1 also shows that the new criterion can perform variable selection in a fashion similar to the Lasso.

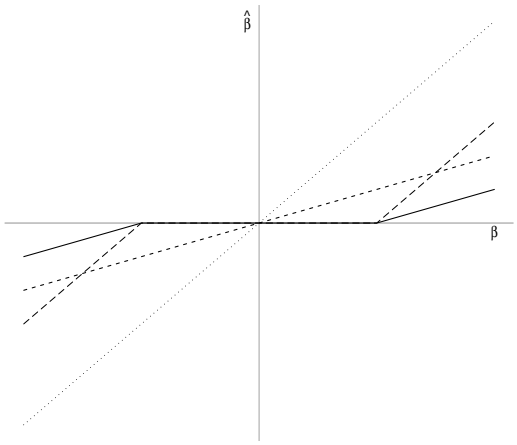


Figure 2: Exact solutions for the Ols (.....), the Lasso (_ _ _ _), CP (.....) and the elastic corr-net (_____) in an orthogonal design: the shrinkage parameters are $\lambda_1 = 2$, $\lambda_2 = 1$ and $p = 2$.

2.3 Orthogonal design case

In the case of an orthogonal design, it is straightforward to show that with parameters (λ_1, λ_2) the elastic corr-net solution is

$$\hat{\beta}_j(\text{elastic corr-net}) = \frac{(|\hat{\beta}_j(\text{ols})| - \lambda_1/2)_+}{1 + 2\lambda_2(p-1)} \text{sign}\{\hat{\beta}_j(\text{ols})\} \tag{2.9}$$

where $\hat{\beta}(\text{ols}) = \mathbf{X}^T\mathbf{y}$ and z_+ denotes the positive part of z, which is z if $z > 0$ and 0 otherwise. The naive elastic-net solution is

$$\hat{\beta}_j(\text{naive elastic-net}) = \frac{(|\hat{\beta}_j(\text{ols})| - \lambda_1/2)_+}{1 + \lambda_2} \text{sign}\{\hat{\beta}_j(\text{ols})\}.$$

The solution of CP regression with parameter λ_2 is given by

$$\hat{\beta}(\text{CP}) = \frac{\hat{\beta}(\text{ols})}{1 + 2\lambda_2(p-1)},$$

which is equal to the ridge regression solution with tuning parameter $2\lambda_2(p-1)$ and the Lasso solution with parameter λ_1 is

$$\hat{\beta}_j(\text{Lasso}) = (|\hat{\beta}_j(\text{ols})| - \lambda_1/2)_+ \text{sign}\{\hat{\beta}_j(\text{ols})\}.$$

It is easy to see that $\hat{\beta}_j$(elastic corr-net) $\leq \hat{\beta}_j$(naïve elastic-net) for all j and all p strictly greater than one.

Fig 2 shows the operational characteristics of the three penalization methods in orthogonal design, where the elastic corr-net is an elastic-net procedure with tuning parameters λ_1 and $2\lambda_2(p-1)$.

2.4 The grouping effect

Qualitatively speaking, a regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated variables tend to be equal (up to a change of sign if negatively correlated). It is too difficult to give in a general case an upper bound, of the absolute difference between any pair (i, j) of the components of $\hat{\beta}$ (the elastic corr-net estimates) as in Zou and Hastie (2005). The following Lemma gives an upper bound of this quantity in the identical correlation case.

Lemma 2.1 (The identical correlation case) *Given data* $(\mathbf{y}, \mathbf{X})$*, where* $\mathbf{X} = (\mathbf{x}_1|...|\mathbf{x}_p)$ *and parameters* (λ_1, λ_2)*, the response is centered and the predictors* $\mathbf{X}$ *standardized. Let* $\hat{\beta}(\lambda_1, \lambda_2)$ *be the elastic corr-net estimate. If* $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$ *and* $\rho_{kl} = \rho$*, for all* (k, l) *, then*

$$\frac{1}{\|\mathbf{y}\|_2}\left|\hat{\beta}_j - \hat{\beta}_i\right| \leq \frac{1-\rho^2}{2(p+\rho-1)\lambda_2}\sqrt{2(1-\rho)}.$$

The proof is differed to an appendix.

Remark 1. In the identical correlation case, the upper bound of the naive elastic-net is

$$\frac{1}{\lambda_2}\sqrt{2(1-\rho)}$$

(Theorem 1 in Zou and Hastie (2005)) which is always greater than the later upper bound of the elastic corr-net. So our method can have potentially a stronger grouping effect in some settings.

For the illustration of the grouping effect we use the idealized example given by Zou and Hastie (2005). With Z_1 and Z_2 being two independent $U(0, 20)$ variables, the response $\mathbf{y}$ is generated as $N(Z_1 + 0.1Z_2, 1)$. Suppose that we observe only

$$\mathbf{x}_1 = Z_1 + \varepsilon_1, \mathbf{x}_2 = -Z_1 + \varepsilon_2, \mathbf{x}_3 = Z_1 + \varepsilon_3$$
$$\mathbf{x}_4 = Z_2 + \varepsilon_4, \mathbf{x}_5 = -Z_2 + \varepsilon_1, \mathbf{x}_6 = Z_2 + \varepsilon_6$$

where ε_i are independent identically distributed $N(0, 1/16)$. 100 observations were generated from this model. The variables $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ may be considered as forming one group and $\mathbf{x}_4$, $\mathbf{x}_5$, $\mathbf{x}_6$ as forming a second group. Fig. 3 compares the solution paths of

the Lasso and the elastic corr-net.

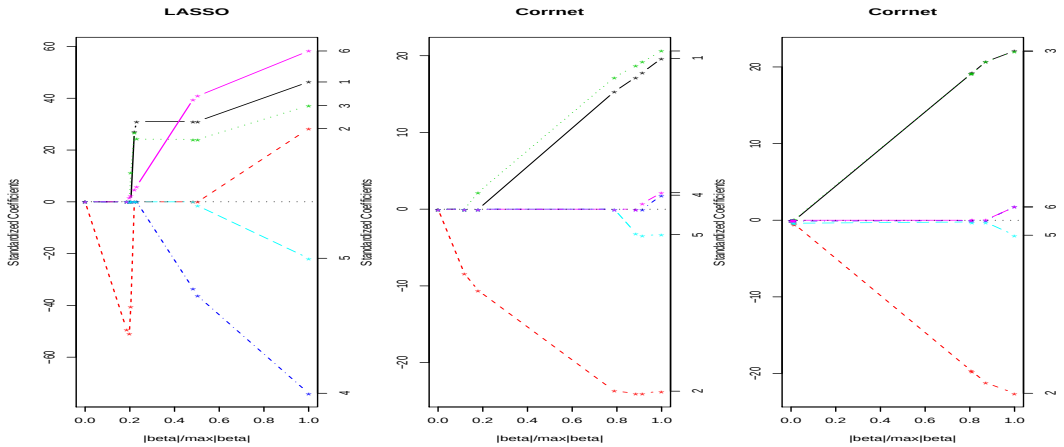


Figure 3: (a) Lasso and (b) elastic corr-net ($\lambda_2 = 0.5$) and (c) elastic corr-net ($\lambda_2 = 1000$) solution paths: the elastic corr-net shows the "grouped selection" $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$ are in one "significant" group and $\mathbf{x}_4, \mathbf{x}_5$ and $\mathbf{x}_6$ are in the other "trivial" group.

3 Computation and tuning parameters selection

3.1 Computation

By Lemma 1, for each fixed λ_2 the elastic corr-net problem is equivalent to a Lasso problem on augmented data set. LARS (Least Angle Regression, Efron et al. 2004) is an efficient algorithm to accelerate the computations of penalized regression parameters as Lasso and related methods. So we use the algorithm LARS to create the *entire elastic corr-net solution path* efficiently with computational efforts of a single OLS fit.

3.2 Tuning parameters selection

In practice, it is important to select appropriate tuning parameters λ_1 and λ_2 in order to obtain a good prediction precision. Choosing the tuning parameters can be done via minimizing an estimate of the out-of-sample prediction error. If a validation set is available, this can be estimated directly. Lacking a validation set one can use ten-fold cross validation. Note that there are two tuning parameters in the elastic corr-net, so we need to cross-validate on a two dimensional surface. Typically we first pick a (relatively

small) grid values for λ_2, say $(0, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 1, 10, 100)$. Then, for each λ_2, LARS algorithm produces the entire solution path of the elastic corr-net. The other tuning parameter is selected by tenfold CV. The chosen λ_2 is the one giving the smallest CV error.

An alternative is to use the uniform design approach of Fang and Wang (1994) to generate candidate points of (λ_1, λ_2). This method actually works for a tuning parameter with arbitrary dimension (cf. Wang et al. 2006). In our experimentations on simulated and real data we use the first approach as in Zou and Hastie (2005).

4 Simulation study

A simulation study was run to examine the performance of the Elastic corr-net, under various conditions with Lasso, Ridge Regression (Ridge), Elastic-net (Enet) and Block-wise boosting (BlockBoost). The simulation setting is similar to the setting used in the original Lasso paper (Tibshirani, 1996) and the elastic-net paper (Zou and Hastie 2005). For each example, the data are simulated from the regression model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

For each example, 50 data sets were generated. Each data set consisted of a training set of size n, on which the model was fitted, an independent validation set of size n is used to select the tuning parameters and a test set is used for evaluation of the performance. In simulations, we centered all variables based on the training data set. Let $\bar{\mathbf{x}}_{\text{train}} = (\bar{\mathbf{x}}_{1,\text{train}}, ..., \bar{\mathbf{x}}_{p,\text{train}})^T$ denote the vector of means of the training data, n_{test} the number of observations in the test data set and $\bar{y}_{\text{train}}$ the mean over the training data.

We use two measures of performance, the test error (mean squared error)

$$\text{MSE}_y = \frac{1}{n_{\text{test}}} \mathbf{r}_{\text{sim}}^T \mathbf{r}_{\text{sim}},$$

estimated on the test data set and the mean squared error for the estimation of β,

$$\text{MSE}_\beta = \|\hat{\beta} - \beta\|_2^2,$$

where

$$r_{i,\text{sim}} = \mathbf{x}_i^T \beta - (\bar{y}_{\text{train}} + (\mathbf{x}_i - \bar{\mathbf{x}}_{\text{train}})^T \hat{\beta}).$$

The four scenarios, considered in Zou and Hastie (2005) and Tutz and Ulbricht (2006), are given by:

1. In Example one, $n = 20$ and there are $p = 8$ predictors. The true parameters are $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma = 3$. with the correlation matrix given by $\rho(\mathbf{x}_i, \mathbf{x}_j) = 0.7^{|i-j|}$. This example contains only positively correlated variables.

2. With $p = 9$, β is specified by $\beta = (1, 2, 3, 4, 0, 1, 2, 3, 4)^T$ and $\sigma = 3$, $\rho(\mathbf{x}_i, \mathbf{x}_j) = 1 - 0.25|i - j|$, the same sample size as in (1). In this example variables are positively and negatively correlated.

3. Example 3 is the same as Example 1, except that $\beta_j = 0.85$ for all j, creating a non-sparse underlying model.

4. In Example 4, $n = 100$ for each of the training and validation sets and there are 40 predictors. The true parameters are

$$\beta = (\underbrace{0, ..., 0}_{10}, \underbrace{2, ..., 2}_{10}, \underbrace{0, ..., 0}_{10}, \underbrace{2, ..., 2}_{10})^T$$

and $\sigma = 15$, with the correlation matrix given by $\rho(\mathbf{x}_i, \mathbf{x}_j) = 0.5$ for $i \neq j$.

Table 1 and Figure 4 summarize both mean squared error of the estimation for the response $\mathbf{y}$ (MSE_y) and the mean squared error for the estimation of β (MSE_β). In Table 1 the best performance is given in boldface. In the four cases, the elastic corr-net outperforms all competitors in both MSE_y and MSE_β followed by the ENET and the RIDGE respectively. The LASSO and BB perform poorly. LASSO is best than BB in examples 1 and 4, while BB is best in examples 2 and 3.

5 High-dimensional experiments

In this section, we give more clarification on the differences between our approach, the Lasso and the elastic net through simulations data in high dimensional setting. Moreover, we examine the performance of the two methods for the identification of relevant variables.

5.1 Predictive power and estimation of effects

In the following, the same notations is used as in the simulations in Section 4. We use the following three high dimensional simulation scenarios correlated groups of variables.

(H1) In this Example the true parameters are

$$\beta = (\underbrace{3, ..., 3}_{15}, \underbrace{0, ..., 0}_{25})^T$$

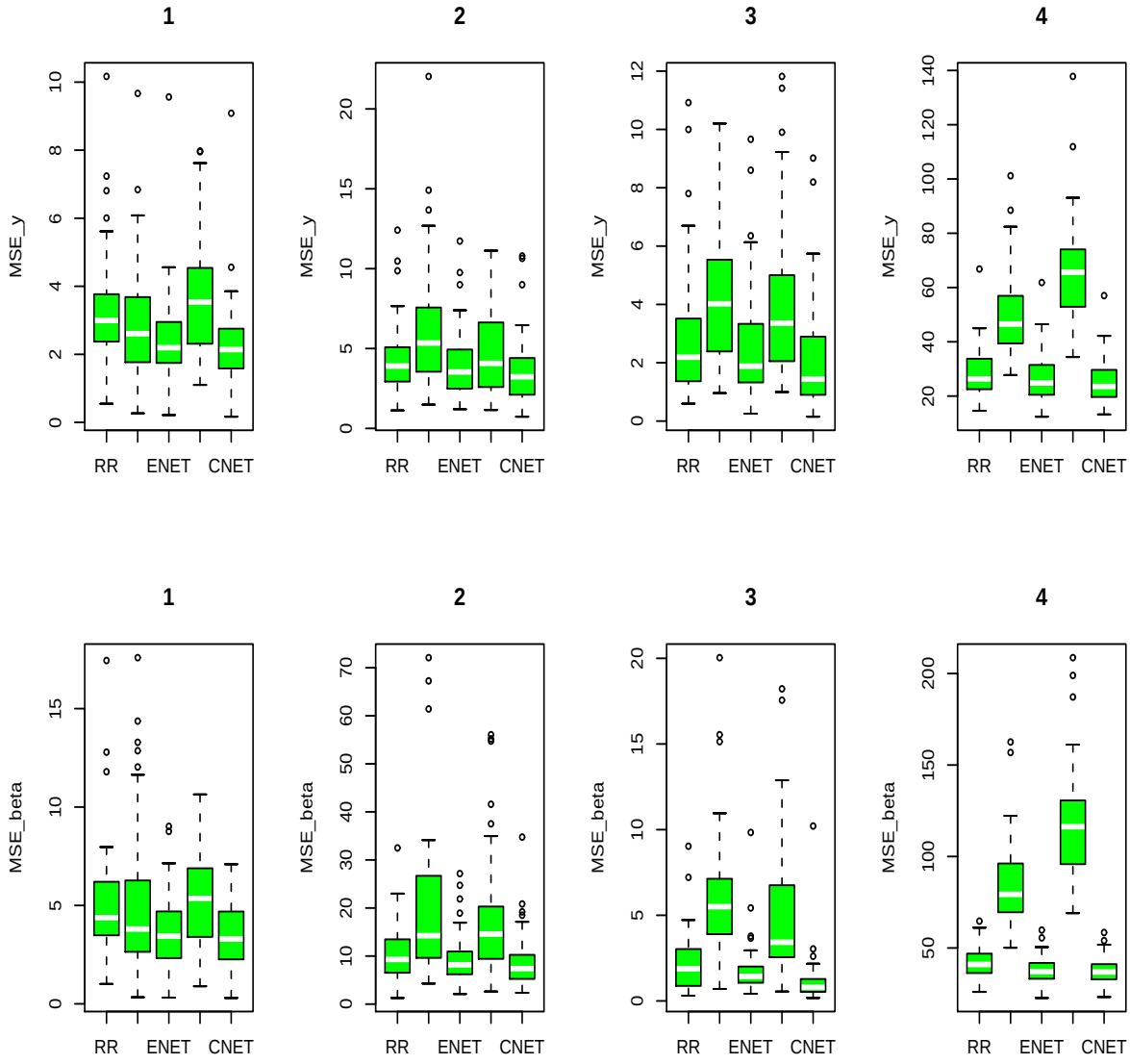
and $\sigma = 15$. The predictors were generated as:

$$\mathbf{x}_i = Z_1 + \varepsilon_i^x, \quad Z_1 \sim N(0, 1), \quad i = 1, ..., 5$$

Example		MSE_y	MSE_β
1	RIDGE	2.99(0.20)	4.38(0.42)
	LASSO	2.60(0.34)	3.80(0.35)
	ENET	2.19(0.16)	3.43(0.32)
	BB	3.53(0.31)	5.36(0.53)
	CNET	**2.14**(0.16)	**3.29**(0.35)
2	RIDGE	3.90(0.31)	9.29(0.84)
	LASSO	5.34(0.46)	14.23(2.27)
	ENET	3.53(0.27)	8.18(0.56)
	BB	4.07(0.46)	14.59(1.64)
	CNET	**3.23**(0.31)	**7.39**(0.97)
3	RIDGE	2.19(0.30)	1.87(0.27)
	LASSO	4.02(0.52)	5.50(0.28)
	ENET	1.88(0.25)	1.43(0.12)
	BB	3.35(0.29)	3.42(0.36)
	CNET	**1.43**(0.26)	**0.80**(0.13)
4	RIDGE	26.33(1.23)	40.99(1.67)
	LASSO	46.55(2.33)	79.32(2.83)
	ENET	24.80(1.22)	37.07 (1.52)
	BB	65.64(2.36)	116.27(5.80)
	CNET	**23.52**(1.16)	**36.89**(1.47)

Table 1: Median mean-squared errors for the simulated examples of five methods based on 50 replications with into parentheses standard errors estimated by using the bootstrap with $B = 500$ resamplings on the 50 mean-squared errors.

Figure 4: Comparing the accuracy of prediction MSE_y and MSE_β of the Ridge (RR), the Lasso, the elastic-net (Enet), the BlockBoost (BB) and the elastic corr-net (C-net) for examples $1-4$.



$$\mathbf{x}_i = Z_2 + \varepsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 6, ..., 10$$
$$\mathbf{x}_i = Z_3 + \varepsilon_i^x, \quad Z_3 \sim N(0,1), \quad i = 11, ..., 15$$
$$\mathbf{x}_i \sim N(0,1), \quad i = 16, ..., 40$$

where ε_i^x are independent identically distributed $N(0,1), i = 1, ..., 15$. In this model the three equally important groups have pairwise correlations $\rho \approx 0.95$, and there are 25 pure noise features. The simulation data has 20/20/40 observations for training set, independent validation set and test set respectively.

(H2) We set $\sigma = 6$ and the true coefficients

$$\beta = (\underbrace{3, ..., 3}_{15}, \underbrace{0, ..., 0}_{25})^T.$$

The predictor were generated as:

$$\mathbf{x}_i = Z_1 + \varepsilon_i^x, \quad Z_1 \sim N(0,1), \quad i = 1, ..., 15$$
$$\mathbf{x}_i \sim N(0,1), \quad i = 16, ..., 40$$

where ε_i^x are independent identically distributed $N(0,1), i = 1, ..., 15$. The simulation data has 20/20/40 observations for training set, independent validation set and test set respectively.

We add a third example considered recently by Whitten and Tishirani (2008),

(H3) Each data set consists of 50/50/400 observations and 50 predictors; $\beta_i = 2$ for $i < 9$ and $\beta_i = 0$ for $i \geq 9$. $\sigma = 2$ and $\rho_{ij} = 0.9 \times 1_{i,j \leq 9}$.

We again measure the performances using the prediction MSE_y and the mean squared error for the estimator of β, MSE_β.

Method	Simulation median MSE_y	H1 median MSE_β	Simulation median MSE_y	H2 median MSE_β	Simulation median	H3 median
Lasso	358.09(32.83)	181.98(16.64)	17.75(1.58)	321.86(24.06)	0.27(0.02)	1.34(0.19)
Enet	151.44(9.86)	74.46(4.07)	**10.90**(1.31)	52.88(7.78)	0.24(0.01)	0.94(0.09)
Corr-net	**138.00**(11.05)	**64.37**(4.26)	11.97(1.34)	**7.20**(1.35)	**0.22**(0.02)	**0.53**(0.05)

Table 2: Median mean-squared errors for the simulated examples and three methods based on 50 replications with standard errors estimated by using the bootstrap with $B = 500$ resamplings on the 50 mean-squared errors.

The simulations show that the elastic corr-net is highly competitive in prediction. Its mean squared error MSE_y is either best or second best in all three examples, while its mean squared error in estimating β (MSE_β) is the best in all examples.

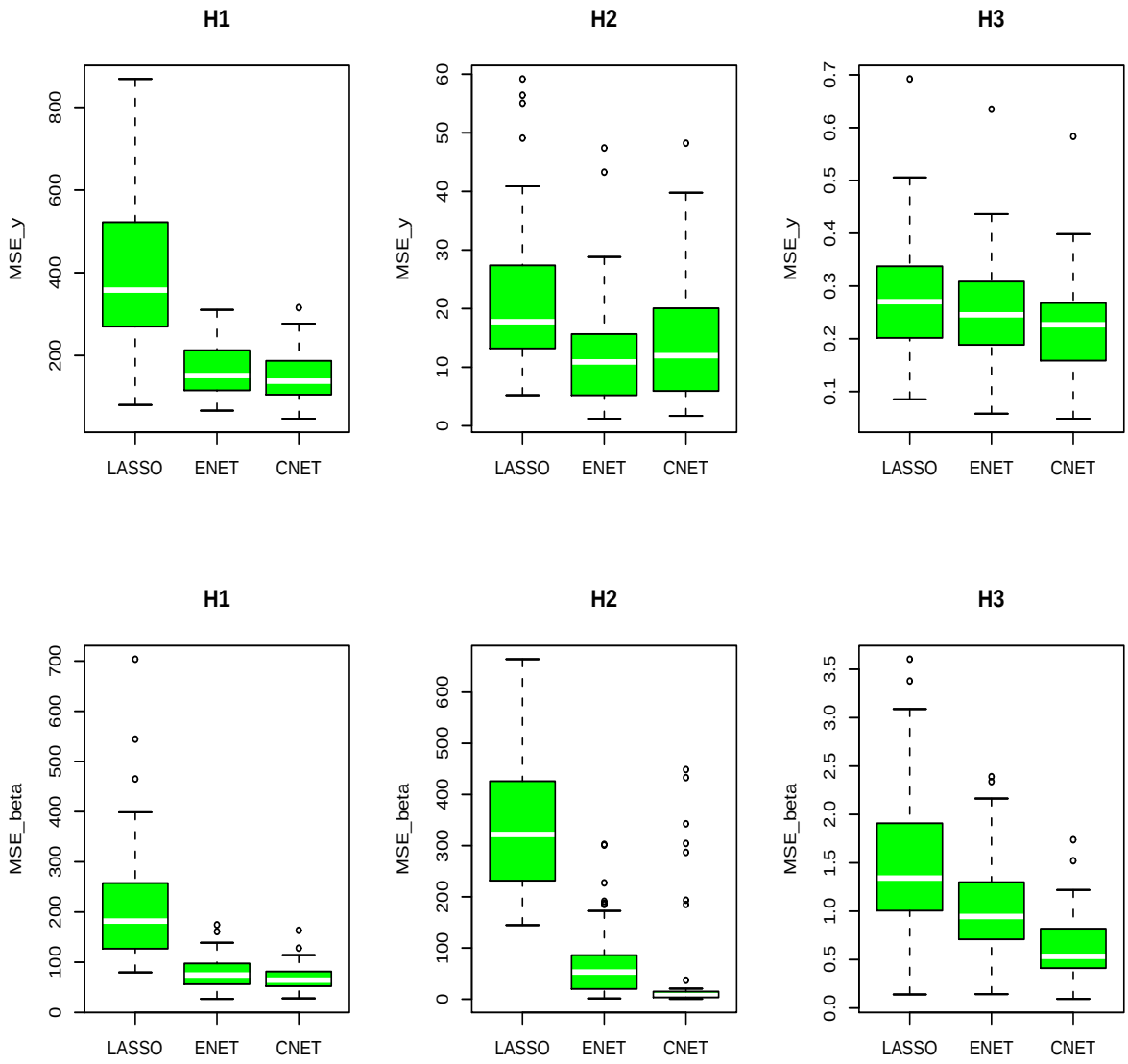


Figure 5: Comparing the accuracy of prediction MSE_y and MSE_β of the the Lasso (LASSO), the elastic-net (ENET) and the elastic corr-net (CNET).

5.2 Identification of relevant variables

There are two fundamental goals in statistical learning: ensuring high prediction accuracy and discovering relevant predictive variables. So the variables included into the

final model are of special interest to practitioners. We want to have a final model as parsimonious as possible but all relevant variables must be present in this model. We can measure the performances of the different methods by the *hits* (the number of selected nonzero components) and the false positives (FP : the number of zero components incorrectly selected into the final model). From the results recorded in Table 3, the Lasso is

Method	Example	H_1	Example	H_2	Example	H_3
	Hits	FP	hits	FP	hits	FP
Lasso	7	5.5	6	3	8	6
Enet	15	16	14	3.5	8	6
Corr-net	15	16	15	4	8	7

Table 3: Median Number of Selected Variables for examples H1, H2 and H3.

not a good variable selection method under collinearity because it eliminates some relevant variables. The elastic corr-net identifies all relevant variables, while the elastic-net has eliminated some one in the second example.

6 Real data sets Experiments

Here we examine the performance of the elastic corr-net for three real world data sets: the body fat data of diemsnsion size $p = 13$, USNews data $p = 13$ and Ionosphere data $p = 33$.

6.1 Analysis the body fat data

The body fat data set has been used by Penrose, Nelson and Fisher (1985). The study aims at the estimation of the percentage of body fat by various body circumference measurements for 252 men. The thirteen regressors are age (1), weight (lbs) (2), height (inches) (3), neck circumference (4), chest circumference (5), abdomen 2 circumference (6), hip circumference (7), thigh circumference (8), knee circumference (9), ankle circumference (10), biceps (extended) circumference (11), forearm circumference (12), and wrist circumference (13). All circumferences are measured in cm. The percent body fat has been calculated from the equation by Siri (1956) using the body density determined by underwater weighting. Figure 6 shows that there are some highly correlated predictors. Some of the pairwise absolute correlation between these covariates are as high as 0.9: weight and hip circumference, chest circumference and abdomen 2 circumference. Seven variables are highly correlated with weight, these variables are: hip circumference, neck circumference, chest circumference, abdomen 2 circumference, thigh circumference, knee circumference and biceps (extended) circumference.

Figure 6: Graphical representation of the correlation matrix of the 13 predictors for the body fat data. The magnitude of each pairwise correlation is represented by a block in the grayscale image.

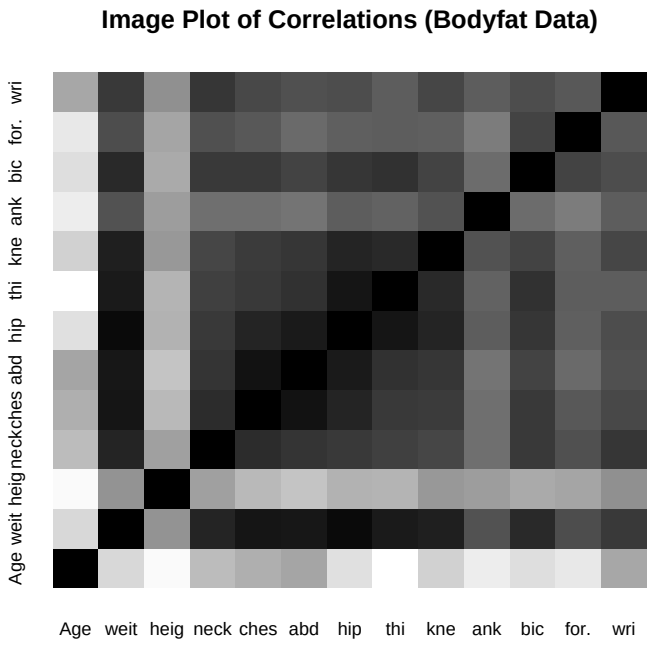


Table 4: Body fat data - median test mean squared error over 20 random splits for different methods.

Method	median MSE_y	median no. of selected variables
Ridge regression	21.02(0.67)	13
Lasso	20.70(1.60)	9.5
Enet	20.23(1.61)	7
BlockBoost	22.01(1.39)	6
Corr-net	**19.77**(1.79)	10

Figure 7: Boxplots of test mean squared errors for 20 random splits of body fat data set into a training set of 151 observations and a test set of 101 observations.

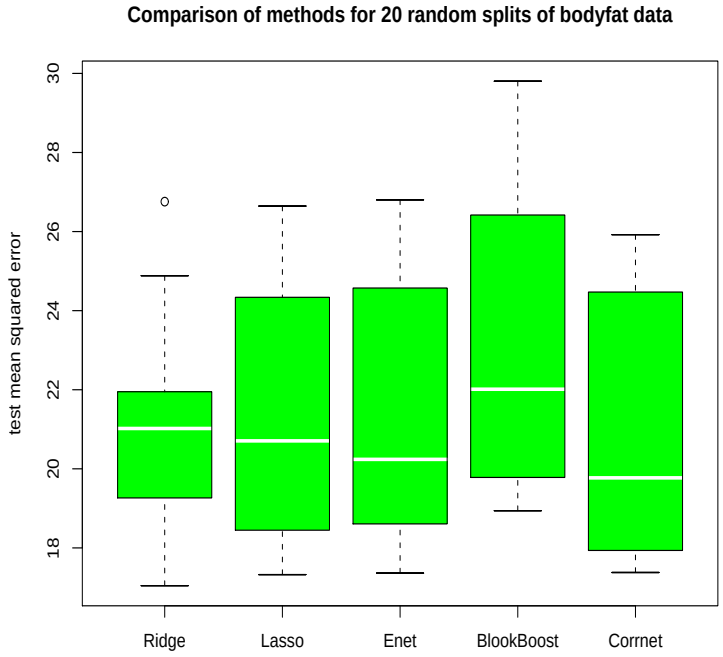


Table 5: Body fat data - tuning parameters and estimated parameters for the whole data set.

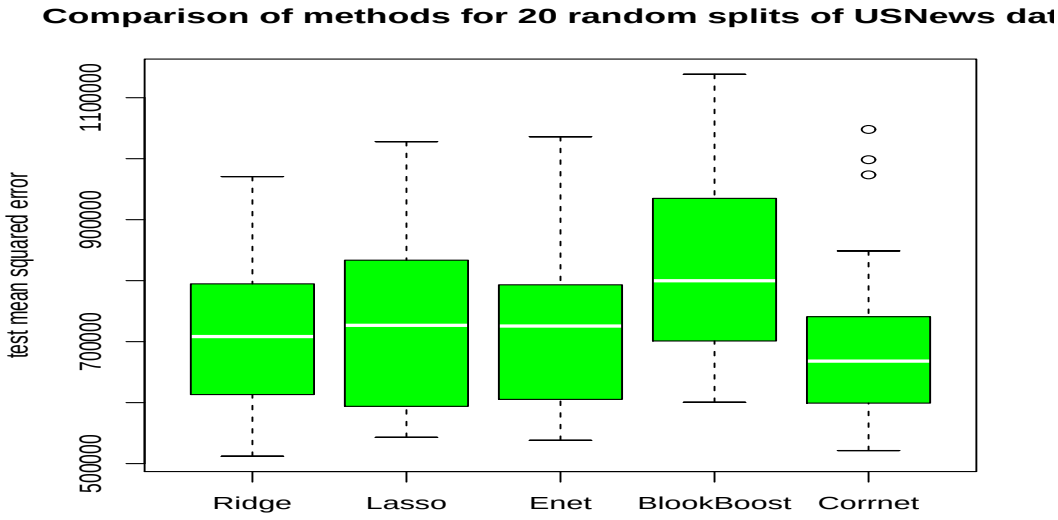
Variables	Ridge	Lasso	Elastic-net	Elastic corr-net
Tuning	$\lambda = 148.41$	$s = 0.79$	$\lambda = 0.05$	$\lambda = 0.05$
parameters:			$s = 0.77$	$s = 0.75$
1	0.07	0.06	0.09	0.06
2	−0.03	−0.05	0.00	−0.02
3	−0.16	−0.11	−0.19	−0.16
4	−0.43	−0.40	−0.24	−0.46
5	0.05	0	0.06	0.00
6	0.77	0.86	0.62	0.86
7	−0.16	−0.11	0.00	−0.18
8	0.19	0.12	0.09	0.08
9	0.10	0	0.00	0.00
10	0.02	0.02	0.00	0.00
11	−0.04	0.10	0.03	0.1
12	0.01	0.37	0.26	0.30
13	−0.39	−1.53	−1.60	−1.61

In order to investigate the performances of the *elastic corr-net*, the data set has been split 20 times into a training set of 151 observations and a test set of 101 observations. Tuning parameter have been chosen by tenfold cross validation. The performance in term of median mean squared error is given in Table 4, the corresponding boxplots are shown in Figure 7. It is seen that the *elastic corr-net* has the best performance in term of mean squared error. While the elastic net seems to miss the relevant variable hip circumference (7),as can be seen from the column of parameter estimation in Table 5.

6.2 USNews data

Our second data set we examine is a subset of USNews data used for the ASA 1995 Data Analysis Exposition. A subset of this data was considered recently by Radchenko and James (2008) (page 16) for styding the performance of their VISA algorithm. The data contains measurements on 18 variables from 777 colleges around the United States. The response of interest is the cost of room and board at each institution. This data contains positively and negatively correlated variables with an average absolute pairwise correlation among the 17 predictors equal to 0.32. We first randomly divide the data into a test data set of 100 observations, with the remainder making up the training data. As for the Body fat data, we have repeated this procedure 20 times. The results are

Figure 8: Boxplots of different methods for 20 random splits of USNews data set into a training set of 100 observations and a test set of 677 observations.



given in Table 6 and Figure 8. The elastic corr-net has the best performances in terms of prediction accuracy.

Table 6: USNews data - median test mean squared error over 20 random splits for different methods.

Method	median MSE_y	median no. of selected variables
Ridge regression	708345.9	18
Lasso	726761.5	8
Enet	725595.0	9
BlockBoost	799844.3	7
Corr-net	**668088.9**	8.5

6.3 Analysis of the Ionosphere data

The Ionosphere data collected by a system in Goose Bay Labrador is considered. The complete data set is of size 351. In this data set, the targets are free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not and their signals pass through the ionosphere (Sigillito et al. 1989). There are 34 continuous predictor variables and the response attribute is either "good" or "bad" radar returns. We exclude the second predictor variable since it takes a constant value 0 for all the observations and we use the remaining 33 to predict the response attribute. This data set was recently used by (Liu and Wu 2007) to illustrate the performance of their regularization regression method based on the combination of the L_0 and L_1 norms.

To explore the performance of different procedures, we randomly divide the data set into three sets of equal sizes for training, tuning, and testing. We repeat this procedure 25 times; the results of variable selection by the Lasso, the elastic-net and the elastic corr-net are given in Figure 9. As shown in the plot, the first, the twentieth and the twenty fifth variables are frequently selected by the three methods. The average rates testing error are 0.0958, 0.0923 and 0.0894 respectively (with their corresponding standard errors 0.0024, 0.0020, and 0.0018 respectively).

7 Discussion

In this paper we have proposed the elastic corr-net for simultaneous estimation and variable selection in linear regression problem. It is a regularization procedure based on the penalized least squares with a mixture of L_1 norm and a weighted L_2 norm penalties. Similar to the elastic-net method, the elastic corr-net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. Additionally, the weighted L_2 penalty explicitly links strength of penalization to the correlation between predictors. Due to the efficient path algorithm (LARS), our procedure enjoys the computational advantage of the elastic-net. Our simulations and empirical results have shown good performance of our method and its superiority over its competitors in term of prediction accuracy, identification of relevant variables while encouraging a grouping effect.

The key contribution of the elastic corr-net is the identification of setting where the elastic net fails to product good results. In fact, our empirical results in term of prediction accuracy show that two setting between the sample size n and the dimension p must be distinguished. First if $p \leq n$, even in small and high dimension settings, the elastic corr-net has shown impressive empirical performance in simulations and world problems. However, in the case of large dimension setting, i. e. $p >> n$, simulations re-

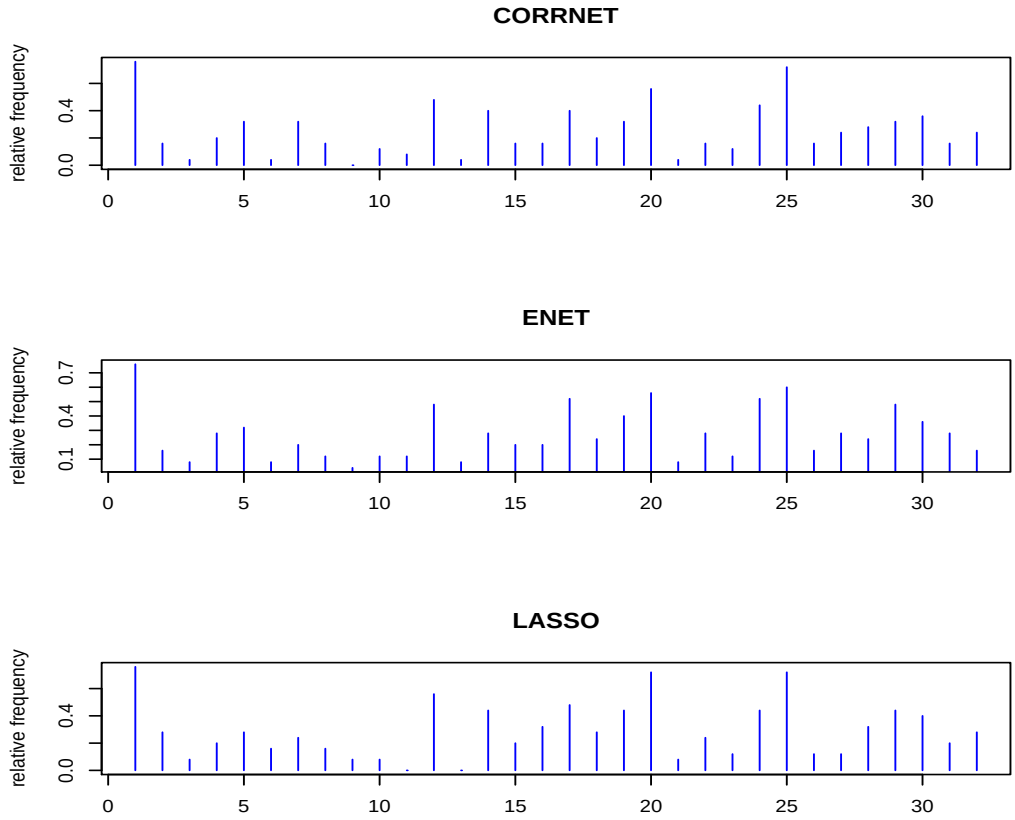


Figure 9: Plot of proportions of the 32 variables selected corresponding to the Lasso, elastic-net and elastic corr-net based on 25 random training samples.

sults suggest very similar performance between the two methods, with a little advantage to the elastic-net.

As the elastic-net, the elastic corr-net can be used in classification problem with the L_2 loss function or the hinge loss function of support vector machine. Its performance with the L_2 loss function seems to be good as shown in section 6.3 with the Ionosphere data. The extension of the elastic corr-net to SVM will be subject to future work.

A Appendix A

A.1 Proof of Lemma 2.1

Proof. Let $\hat{\beta}(\lambda_1, \lambda_2) = \arg\min_\beta\{L(\lambda_1, \lambda_2, \beta)\}$. If $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$, then both $\hat{\beta}_i(\lambda_1, \lambda_2)$ and $\hat{\beta}_j(\lambda_1, \lambda_2)$ are non-zero, and we have $\text{sign}(\hat{\beta}_i(\lambda_1, \lambda_2)) = \text{sign}(\hat{\beta}_j(\lambda_1, \lambda_2))$. Then $\hat{\beta}(\lambda_1, \lambda_2)$ must satisfies

$$\frac{\partial L(\lambda_1, \lambda_2, \beta)}{\partial \beta}|_{\beta=\hat{\beta}(\lambda_1,\lambda_2)} = \mathbf{0}. \tag{A.10}$$

Hence we have

$$-2\mathbf{x}_i^T\{\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_1, \lambda_2)\} + \lambda_1 \text{sign}\{\hat{\beta}_i(\lambda_1, \lambda_2)\} + 2\lambda_2 \sum_{k=1}^{p} \omega_{ik}\hat{\beta}_k(\lambda_1, \lambda_2) = 0, \tag{A.11}$$

$$-2\mathbf{x}_j^T\{\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_1, \lambda_2)\} + \lambda_1 \text{sign}\{\hat{\beta}_j(\lambda_1, \lambda_2)\} + 2\lambda_2 \sum_{k=1}^{p} \omega_{jk}\hat{\beta}_k(\lambda_1, \lambda_2) = 0, \tag{A.12}$$

Subtracting equation (A.11) from (A.12) gives

$$(\mathbf{x}_j^T - \mathbf{x}_i^T)\{\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_1, \lambda_2)\} + \lambda_2 \sum_{k=1}^{p} (\omega_{ik} - \omega_{jk})\hat{\beta}_k(\lambda_1, \lambda_2) = 0,$$

which is equivalent to

$$\sum_{k=1}^{p} (\omega_{ik} - \omega_{jk})\hat{\beta}_k(\lambda_1, \lambda_2) = \frac{1}{\lambda_2}(\mathbf{x}_i^T - \mathbf{x}_j^T)\hat{\mathbf{r}}(\lambda_1, \lambda_2), \tag{A.13}$$

where $\hat{\mathbf{r}}(\lambda_1, \lambda_2) = \mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_1, \lambda_2)$ is the residual vector. Since $\mathbf{X}$ is standardized, then

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = 2(1 - \rho_{ij}).$$

Because $\hat{\beta}(\lambda_1, \lambda_2)$ is the minimizer we must have

$$L\{\lambda_1, \lambda_2, \hat{\beta}(\lambda_1, \lambda_2)\} \leq L\{\lambda_1, \lambda_2, \beta = \mathbf{0}\},$$

i.e.

$$\|\hat{\mathbf{r}}(\lambda_1,\lambda_2)\|^2+\lambda_2\hat{\beta}^T(\lambda_1,\lambda_2)\mathbf{W}\hat{\beta}(\lambda_1,\lambda_2)+\lambda_1\|\hat{\beta}(\lambda_1,\lambda_2)\|_1\leq\|\mathbf{y}\|_2^2.$$

So $\|\hat{\mathbf{r}}(\lambda_1,\lambda_2)\|_2\leq\|\mathbf{y}\|_2$. Then the equation (A.13) implies that

$$\frac{1}{\|\mathbf{y}\|_2}\mid\sum_{k=1}^{p}(\omega_{ik}-\omega_{jk})\hat{\beta}_k(\lambda_1,\lambda_2)\mid\leq\frac{1}{\lambda_2\|\mathbf{y}\|_2}\|\hat{\mathbf{r}}(\lambda_1,\lambda_2)\|\|\mathbf{x}_1-\mathbf{x}_2\|\leq\frac{1}{\lambda_2}\sqrt{2(1-\rho_{ij})}\quad\text{(A.14)}$$

We have:

$$\omega_{ii}=-\sum_{s\neq i}\frac{\omega_{is}}{\rho_{is}}\qquad\text{and}\qquad\omega_{jj}=-\sum_{s\neq j}\frac{\omega_{js}}{\rho_{js}}.\qquad\text{(A.15)}$$

Then

$$\sum_{k=1}^{p}(\omega_{ik}-\omega_{jk})\hat{\beta}_k(\lambda_1,\lambda_2)=\frac{-2}{1-\rho_{ij}}[\hat{\beta}_j(\lambda_1,\lambda_2)-\hat{\beta}_i(\lambda_1,\lambda_2)]+2SN\qquad\text{(A.16)}$$

where

$$SN=\sum_{k\neq i,j}\frac{1}{1-\rho_{ki}^2}[\hat{\beta}_i(\lambda_1,\lambda_2)-\rho_{ki}\hat{\beta}_k(\lambda_1,\lambda_2)]+\frac{1}{1-\rho_{kj}^2}[\rho_{kj}\hat{\beta}_k(\lambda_1,\lambda_2)-\hat{\beta}_j(\lambda_1,\lambda_2)].$$

if

$$\rho_{ki}=\rho_{kj}=\rho,\quad\forall k=1,...,p,$$

we have

$$SN=\frac{p-2}{1-\rho^2}(\hat{\beta}_i(\lambda_1,\lambda_2)-\hat{\beta}_j(\lambda_1,\lambda_2)).$$

So using (A.14) we have:

$$\begin{aligned}\frac{1}{\|\mathbf{y}\|_2}\left|\hat{\beta}_j-\hat{\beta}_i\right| &\leq \frac{1-\rho^2}{2(p+\rho-1)\lambda_2}\sqrt{2(1-\rho)}\\ &\leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)}.\end{aligned}$$

This completes the proof ■

Acknowledgements

The authors would like to thank Gilles Celeux and Jean Michel Marin for valuable comments of the first versions of this paper.

References

[1] Bondell, H. D. and Reich, B. J. (2007). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics* **64**, 115 − 123..

[2] Chen, S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit, *SIAM J. on Sci. Comp.,* **20**, no. 1, 33 − 61,

[3] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics,* **32**, 407 − 499.

[4] Fang, K.-T. and Wang, Y.(1994). Number-Theoretic Methods in Statistics. Chapman and Hall: London.

[5] Hoerl, A. and Kennard, R.(1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics,* **12**, 55 − 67.

[6] Liu, Y. and Wu, Y.(2007). Variable selection via a combination of the L_0 and L_1 penalties., *Journal of Computational and Graphical Statistics*, **16**, **4**, 782 − 798.

[7] Penrose, K. W., Nelson, A. G., and Fisher, A. G. (1985). Generalized body composition prediction equation for men using simple measurement techniques. *Medicine and Science in Sports and Exercise* **17**, 189.

[8] Radchenko, P. and James, G. M.(2008). Variable inclusion and shrinkage algorithms, *Journal of the americain statistical association (To appear).*

[9] Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989), Classification of Radar Returns from the Ionosphere Using Neural Networks, *Johns Hopkins APL Technical Digest,* **10**, 262 − 266.

[10] Siri, W. B.(1956). The gross composition of the body. In C. A. Tobias and J. H.Lawrence (Eds.), *Advances in Biological and Medical Physics*, Volume 4, pp. 239 − 280. Academic Press New York.

[11] Tibshirani, R.(1996). Regression shrinkage and selection via the Lasso, *ournal of the Royal statistical Society, B.* **58**, 267 − 288.

[12] Tutz, G. and Ulbricht, J. (2006). Penalized regression with correlation based penalty. Discussion Paper 486, SFB 386, Universität München.

[13] Zou, H. and Hastie, T. (2005)., Regularization and variable selection via the elastic-net, *Journal of the Royal statistical Society, B.* **67**, 301 − 320.

[14] Wang, S., Nan, B. Zhu, J. and Beer, J. (2006), Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics.* (To appear.)

[15] Wang, L., Zhu, J. and Zou, H.(2006), The Doubly Regularized Support Vector Machine. *Statistica Sinica,* Vol. **16**(**2**), 589 – 616.

[16] Witten, D. M. and Tibshirani, R. (2008), Covariance-regularized regression and and classification for high-dimensional problems. *Journal of Royal Statistical Society*, Series **B**, to appear.

Contents

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399